# DISCOVERING COMPLEX RELATIONSHIPS BETWEEN DRUGS AND DISEASES

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Ranjana Sharma

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

November 2011

Fargo, North Dakota

# North Dakota State University

Graduate School

**Title**

DISCOVERING COMPLEX RELATIONSHIPS BETWEEN DRUGS AND DISEASES

**By**

RANJANA SHARMA

The Supervisory Committee certifies that this *disquisition* complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Juan (Jen) Li
Chair

Dr. Kendall Nygard

Dr. Gursimran Walia

Dr. Ross Collins

Approved:

| 11/28/2011 | Dr. Brian M. Slator |
|---|---|
| Date | Department Chair |

ABSTRACT

Finding the complex semantic relations between existing drugs and new diseases will help in the drug development in a new way. Most of the drugs which have found new uses have been discovered due to serendipity. Hence, the prediction of the uses of drugs for more than one disease should be done in a systematic way by studying the semantic relations between the drugs and diseases and also the other entities involved in the relations. Hence, in order to study the complex semantic relations between drugs and diseases an application was developed that integrates the heterogeneous data in different formats from different public databases which are available online. A high level ontology was also developed to integrate the data and only the fields required for the current study were used. The data was collected from different data sources such as DrugBank, UniProt/SwissProt, GeneCards and OMIM. Most of these data sources are the standard data sources and have been used by National Committee of Biotechnology Information of Nation Institute of Health. The data was parsed and integrated and complex semantic relations were discovered. This is a simple and novel effort which may find uses in development of new drug targets and polypharmacology.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Juan (Jen) Li, for her excellent guidance, caring and patience. She provided me encouragement and helped me even when she herself was going through a tough time.

I would like to thank Dr. Kendall Nygard, who welcomed me in the Department and provided all the support whenever I needed it. Special thanks go to Dr. Nygard, Dr. Gursimran Walia and Dr. Ross Collins, who were willing to participate in my final defense committee at the last moment.

I would also like to thank Prof. M.P.Singh, Ex-Director and Prof. H.S.Saxena Ex-Dy Director of AIT for allowing me study leave to pursue my degree.

My sincere thanks to my good friend Benjamin Bengfort, for helping me with the implementation of the project. He was always willing to give his best suggestions and support. Many thanks to Chainika Jangu, Navneet Deosi, Ankit Kumar, Shweta Tiwari, Richa Gautam, Kunjan Gautam, Bithika Sharma and Sanjay Sharma for providing me all the fun times.

I would like to thank my parents, and my husband, Arun, for supporting me in my decision to pursue my degree and allowing me to come to a different country although they themselves were lonely. I will always be grateful to them for this gesture.

Last but not the least I would like to thank my son Akshat and Karan for standing by me and providing me all the support whenever I needed it.

(RANJANA SHARMA)

TABLE OF CONTENTS

ABSTRACT ..... iii

ACKNOWLEDGEMENTS ..... iv

LIST OF TABLES ..... vii

LIST OF FIGURES ..... viii

LIST OF ALGORITHMS ..... ix

CHAPTER 1. INTRODUCTION ..... 1

CHAPTER 2. RELATED WORK AND BACKGROUND KNOWLEDGE ..... 7

2.1 Semantic Web ..... 7

2.2 RDF (Resource Description Framework) ..... 8

2.3 Generic Ontological Triples ..... 9

2.4 Semantic Relations ..... 9

2.5 Ranking Semantic Associations ..... 11

2.6 Bio2RDF: Linked Data for the Life Sciences ..... 14

2.7 Linked Open Drug Data ..... 14

2.8 Drug Delivery and Polypharmacology ..... 15

2.9 Polypharmacology or Drug Repositioning ..... 16

CHAPTER 3. SYSTEM DESIGN ..... 18

3.1 Overview ..... 18

3.2 Data Sources ..... 20

3.2.1 Integration of Data Sources ..... 21

3.3 Methods of Complex Relation Discovery ..... 21

3.3.1 Algorithm ..... 21

3.4 Problem Statement ..... 23

3.5 Contribution to the Paper ..... 23

3.6 Experimental ..... 24

3.6.1 Omim Model ..... 24

3.6.2 Software Requirements ..... 24

3.6.3 Data Sources ..... 25

3.6.4 Ontological Management ..... 30

CHAPTER 4. CONCLUSION AND FUTURE WORK ..... 39

4.1 Conclusion ..... 39

4.2 Future Work ..... 40

REFERENCES ..... 42

LIST OF TABLES

Table Page

3.1.Sample Data from DrugBank.....26

3.2.Sample Data from www.uniprot.org.....27

3.3.Sample Data from www.genecards.org.....28

3.4.Relationship between the Drug and Disease through MIM#.....34

3.5.Relationship between Drug and Disease without the Use of MIM#.....34

3.6.Relationship between Drug and Disease Directly through Protein.....35

LIST OF FIGURES

Figure Page

1.1.Relation between Two Entities through Different Paths....4

2.1.The Process of Drug Development....16

3.1.The Omim Component Architecture....25

3.2.Schema....29

3.3.Screenshot of the Ontology from Protégé....31

3.4.More than 1 Path from Drug to Disease....36

3.5.The Drug-Disease-Gene Relationship for Gene ACHE....37

LIST OF ALGORITHMS

Algorithm Page

1. Iterative Deepening Algorithm for Relation Discovery…….22

# CHAPTER 1. INTRODUCTION

Drug discovery is a process of discovering and designing drugs and is generally related to the fields of medicine, pharmacology and biotechnology. Despite advances in technology and understanding of biological systems, drug discovery is still a lengthy (taking about 15 years of time from proposal to being used as treatment for patients [1]), expensive (costing about $800 million to $1 billion [1]), difficult, and inefficient process with low rate of new therapeutic discovery [1]. Development of a new drug from scratch is a complex process that involves extensive research of the proposed compounds. Between 5000 to 10,000 compounds that are proposed for a potential drug, only about 2.5% to 5% of the compounds get approved for preclinical trial. Out of these only 2% may get approved for clinical trial, further only 1 compound becomes an approved drug [1 - 4].

Drug repositioning is an effective solution to the aforementioned problem [5]. It is the application of the existing drugs to new indications or new diseases. An existing drug has passed significant pre-clinical and clinical tests. Its toxicity and other effects are already known. Hence the cost of using it for some other disease will be much less compared to developing a drug from scratch. Drug repositioning is the study of interaction of drugs with multiple targets. Conventional drug design follows the principle of “one gene, one disease, one drug” i.e. one drug is targeted for the treatment of one disease caused by one gene. However, the same drug can be used with multiple diseases. Most of the repositioned drugs that are being currently used have been developed by chance on observation of the side-effects of the drug [6,7]. A compound or a drug may be related to new diseases. For example the side effect of one drug may be the treatment of another disease. However, this relationship might not have been identified earlier and the

information relating the drug and the new application might not be available in detail. Hence, a detailed study involving the proteins, genes, pathway and other important factors should be carried out to study the polypharmacological action of drugs rather than discovering the effects by mere observation.

To solve the aforementioned problems, the main aim was to discover valuable relationships between drugs, proteins, genes, and diseases for this study. The present work involves the integration of data over various diverse domains such as chemical, biological etc. The data that has to be integrated over these domains may also be disparate data in multidimensional and heterogeneous forms. To reconcile the heterogeneity problem, Semantic Web and knowledge representation techniques have been adopted for effective integration of data as well as find the appropriate associations between the drugs and diseases and therefore find the most effective drug against a disease by providing appropriate ranking of the drugs.

In the proposed approach, Semantic Web techniques are used to represent the data sources used in the study. Semantic Web is the extension of the current World Wide Web which contains a large amount of linked information over various platforms. Combining and linking such disparate data is a challenge today which can be assisted with Semantic Web technologies. The focus in this work, in particular, is on Semantic Associations. Semantic Associations are complex relationships between entities, events and concepts. With the help of these associations the information contained can be understood and interpreted differently. They may even provide information that may not be easily determined otherwise.

An entity may be related to another entity directly or through one or more intermediate entities. An association where an entity is related to another entity through intermediate entities is extremely important in various applications. In the human body a number of processes take place. The chemicals in the body may be broken down to make simpler chemicals or form other chemicals used in the functioning of the body. This involves various genes, proteins, enzymes etc. When one molecule is transformed into another molecule through a series of steps, the process is called as pathway. The pathway is catalyzed by enzymes at various steps. Diseases affect the body physically and may affect the internal processes of the body and various pathways. In order to cure the diseases, drugs are used which are foreign molecules. These foreign molecules are converted into other molecules and help in curing the disease. This might involve inhibiting the pathway of various organisms including bacteria causing the disease. Thus the drugs may be related to diseases either through the protein targets or through the genes and pathways. The information about the various molecules in the body and their properties and functions as well as the processes that interact with these molecules is widely available in the public databases such as DrugBank [8], KEGG [9-11], OMIM [12]. The drugs can be related to the diseases via intermediate protein targets. The protein targets are directly related to the diseases through OMIM database [10]. The drugs and the diseases may also be related via genes and pathway they inhibit.

An entity may be related to another entity in one or more than one way. The entity e1 and e5 could be related through different entities such as e2, e3 or e4. The different paths that may exist between e1 and e5 could be e1 – e2 – e5, e1 – e3 – e5 or e1 – e4 – e5 (Figure 1.1). For example a student is related to Professor by the relationship “teaches” but

may also be related through other ways as both student and Professor may be residing in the same geographical location or they may be the member of same organization, or they may even have same hobbies and interests. A drug can be linked to a disease by more than one protein target as well as more than one side effects that it may have on the human being.

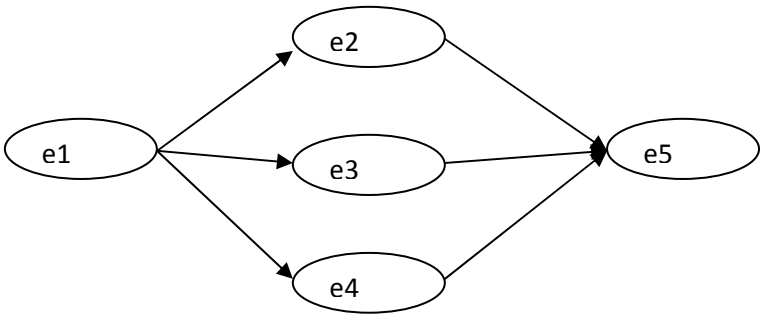


Figure 1.1. Relation between Two Entities through Different Paths

Thus there may be hundreds of such associations between two entities. However, many of these paths may not provide meaningful information to the user and may be considered as irrelevant as the user may not consider the context in which those paths could be useful. Hence, in order to find the relevant paths or Semantic Associations, the user's domain or region of interest or the context in which the associations would be interpreted should be determined.

A Semantic Association is a sequence of complex relationship between the entities which may be from disparate sources. They have a directed path from one entity to another. Also, more than one path may exist between the entities or they may be associated in more than one way. Hence, a search algorithm is needed to return one or more paths result. In order to do that the search algorithm needed to be constrained (bounded) so that

the results are meaningful. In the approach presented here, the associations between the drugs and diseases were considered. These associations were found by integrating the data of drugs, proteins, targets, genes and diseases across several websites. Once the associations were found, they were examined for the number of ways in which one entity is associated with another. As stated earlier, the search can result into a number of paths which may not be useful to the user as they may not be in context in which the user is querying. Hence, these paths need to be relevant and the relevance could be found by ranking the paths in between the Semantic Associations.

The association between two entities within the domain of interest or context can be ranked based on their relevance. An association is ranked the highest when it provides the maximum information in a given context. Hence, the context should be defined within which the associations need to be ranked. The context also includes the associated properties and the classes. Besides this the length of the association is also considered important to provide the relevant information

It has been observed that there is no single method for ranking the Semantic relationships between the entities, as a number of factors have to be considered. Hence, an approach had to be considered that had an edge over the other algorithms reported. Also the method should be flexible enough to consider the different factors involved such as length of the association, context in which it is to be determined, as well as the frequency of their occurrence.

Hence in the present work, the iterative deepening algorithm has been used to search for the paths in Semantic Associations between the drugs and the diseases. The

associations between the drugs and the diseases are identified using intermediary genes, proteins and targets. The algorithm also considers many factors that may be involved during the ranking approach. As stated earlier, the drug development is a lengthy and expensive process that may not give very promising results. Hence, the new uses for existing drugs would be more useful than developing a drug from scratch. Rather than finding the new uses by chance, studies could be carried out for finding other uses. In this work, an attempt of a systematic study between the drugs and their effects on the diseases and the side effects that may lead to the new use of existing drugs was made.

In summary, the contribution of this paper can be summarized as:

1. A novel scheme is proposed to discover drugs by mining data from dispersed drug-related knowledgebases.

2. The present application helps in finding the Semantic Associations between the drugs and the diseases from the ontology prepared using disparate data from different websites.

The following of the paper is organized as follows. Chapter 2 provides an overview of some background knowledge and related work, Chapter 3 gives the System Design as to how the studies on finding the appropriate paths or Semantic Associations between the drug and the disease can be found out. It also gives results and discussion and Chapter 4 provides the conclusions and future work. The References used in the paper are listed in the end.

# CHAPTER 2. RELATED WORK AND BACKGROUND KNOWLEDGE

## 2.1 Semantic Web

Semantic Web can be considered as the largest repository of information on the web. The information contained on the Semantic Web has the user terminology, vocabulary and the language from across the globe. It provides the capability of storing multiple names of the same product or aliases, with the help of which the users of Semantic Web can search using different name, spellings of that name, code names, or acronyms for the same product as well as producing the same result upon querying by linking and integrating the disparate data[13]. Semantic Web can interpret web pages consisting of knowledge and data that can be processed by computers making it much more understandable without the human intervention

The Semantic Web defines data on the Internet in a meaningful and functional way that is specifically understandable by computers. Internet information retrieval systems (search engines) will soon be able to more efficiently query and retrieve relevant information from documents by extracting semantic metadata from document headers. However, semantic data is able to reveal much more complex relationships than simple information retrieval is able. By querying semantic relationships artificially intelligent agents are able to discover complex relationships by searching for paths via chaining semantic triples.

As more and more web bases store semantic metadata on their pages, the number of relationships between entities on the Internet, and descriptions of data will need to be further defined. Basic ontological structures are being developed to constrain that information, but only in schema form. Instances of semantic structures still reside in

text/xml format on the pages, and in order to be searched efficiently they must be managed in a more structured system. In addition, as the number semantic relationships explodes on the Internet, efficient search and ranking algorithms will need to be created in order to digest the data and discover complex semantic relationships- i.e. relationships between web entities that span across several semantic hops.

The Semantic Web itself is a globally linked mesh of ontological information to be understood by machines. This, in turn, is a subset of the World Wide Web which is a globally linked mesh of information understood by humans. The Semantic Web makes use of ontological schemas to define data in web documents and creates instances of those objects defined in the schemas as metadata added to web documents. This metadata can then be searched by artificially intelligent agents to produce meaningful search results or provide more detailed about web resources.

## 2.2 RDF (Resource Description Framework)

Generally, XML RDF is the format used for the semantic metadata added to web pages- and is quickly becoming a W3C standard for this purpose. RDF provides an ontological description of a resource and is in the basic format of a generic ontological triple (discussed in the next section). Representing the triples efficiently, is not as easy as it sounds, different types of formats have been developed, for external and internal processing purposes. One of these external RDF formats is known as Notation3, a plain text format devised by Tim Berners-Lee (who proposed World Wide Web and Semantic Web), which is easy to learn, and easy to process [14].

## 2.3 Generic Ontological Triples

A triple is a semantic statement that describes a relationship, and is usually defined using RDF and RDF Schemas [14]. They define properties of the resource they describe or create relationships between different resources. Their basic purpose is to make natural language data "understandable" to machines, so that relationships or objects can be processed for various purposes instead of just parsed. Triples are built on two ontological structures: entities and predicates, where entities are the subject and object of a predicate-which is a relationship between the two entities [15]. They usually take the form: <Subject, Predicate, Object>. This triple can be read: "The subject entity is related to the object entity by the predicate", or alternatively "The subject predicates the object".

Usually the two entities in a triple have URIs (Uniform Resource Identifiers), in order to maintain the universality of those objects. Using URIs allows us to know the exact resource of the statement and the property to be assigned to it. Triples may consist of three URIs, one for both entities and the predicate have the strongest description. The second generation of the Semantic Web will focus on adding semantic annotations that software can understand and from which humans can also benefit because of its natural readability – but it will always necessarily be in triple form.

## 2.4 Semantic Relations

The main goal of this paper is to discover complex semantic relationships between entities defined by ontologies. A complex semantic relation can be defined as the path between two entities where a path is defined as:

$$e_1, P_1, e_2, P_2, \dots, e_{n-1}, P_{n-1}, e_n$$

where $e_i$ is an entity and $P_j$ is a predicate or property that defines a relationship between two entities, $e_i$ and $e_{i+1}$. Therefore this denotes a complex semantic relation between entities $e_1$ and $e_n$. Essentially this is the chaining of ontological triples, where the next triple has subject that was the object of the previous triple:

<$Subject_1$, $Predicate_1$, $Object_1$> <$Object_1$, $Predicate_2$, $Object_2$>…<$Object_{n-1}$,$Predicate_n$, $Object_n$>

Semantic associations are meaningful complex relationships between entities, events and concepts. They provide meaningful information so that it is understandable and new relationships could be discovered over the Semantic Web [16]. Various studies have been carried out over the years to find the complex relationships between the entities using Semantic Web. Complex semantic relationships between the entities have a directed path or a sequence to get from one entity to another [17]. RDF is a standard for describing and exchanging the information of the web resources. Although querying is not well supported in RDF, ρ-queries help in querying for Semantic Associations on the Semantic Web and find some relationships [17].

An important measurement for these complex relations is the number of hops it takes to reach $e_n$ from $e_1$. Hop count is the number of traversals down a directed acyclical graph tree created by these relationships. Alternatively, the number of hops is the number of triples required to generate the path. Hops are also a measure of depth for these relationships.

Because of the extremely large metabase being queried, there may be too many results many of which may not be of importance. Therefore ranking of complex

relationships discovered is also of vital importance. There is a need for a flexible ranking approach that would allow the identification of the most interesting and meaningful Semantic Associations between two entities which is effective and efficient in terms of time, space and relationship complexity. Unfortunately there is no natural heuristic for determining this since there is no cost other than hops associated with these relations.

## 2.5 Ranking Semantic Associations

Searching and ranking of documents can be done within some relevance measures such as relevance weights that are assigned to the relationships by the humans and a relevance threshold. The relationships of named entities can be analyzed with respect to a query. The results can be obtained within this threshold [18-20].

The research in life sciences had produced huge amount of data. Large amount of this data is available as public databases for further research and studies. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. [21]. Open life science and biomedical Ontologies are present on the World Wide Web. Since these ontologies are from varied sources they can be associated using RDF as triples. Hence, molecular biology could be related to bioinformatics which in turn could be related to drug delivery by relating specific gene to a specific protein which could be a target protein or a specific drug. The genes and the proteins could be related using the SwissProtID. The associations between drugs and the metabolic pathways could also be studied using the Semantic Web. The genes are associated with the metabolic pathways using the PathwayID [22]. Thus, gene data from a genomic database and corresponding protein data from a protein database can be merged together within RDF documents giving meaningful associations.

Also, since many association paths can exist between any two given entities, these paths are needed to be ranked to find the most relevant path within the system biology space. The entities can be associated with each other based on the context in which they are being considered. The genes can be easily associated with the proteins in the similar way If the proteins are localized in a particular tissue (context) then those proteins could be linked to the corresponding genes by the functional analysis of the genes.

Semantic Web ranks different types of entities on the web. This includes the documents, relationships, queries, ontologies etc. Different approaches have been employed to determine the ranking of each of these. Ranking of the semantic relationships is different than ranking of documents. In Semantic Web ranking the queries are ranked by finding the relevance in the keywords returned as the result of the query. The results of the query should be the association between entities which is the path between them. A very popular algorithm, PageRank Algorithm is a link analysis algorithm which is commonly employed to find the ranking of the documents hyperlinked to each other by assigning the weights to the incoming as well as outgoing links. It uses a wide data with 500 million variables and 2 billion terms [23]. The ObjectRank system applies authority-based ranking to keyword search in databases modeled as labeled graphs [24]. The nodes contain the keywords and are connected semantically to the other nodes by authority. Each node is ranked according to its authority with respect to the particular keywords. Weights are assigned to global importance, each keyword of the query as well as the importance of a result actually containing the keywords. The results that actually contain the keywords are ranked higher than being referenced by nodes containing them.. Teoma is a search engine which is used or ranking the search results on the web. It analyzes the web pages and

groups them based on their subject [25]. This algorithm also uses the popularity technique to rank the results. ReConRank [26] is a graph based ranking algorithm. The ranking is applied in the depth to graphs and subgraphs. It combines two approaches. It ranks the RDF graphs by ranking the data graph as well as the data provenance graphs. SemRank [27] is an algorithm based on relevance and ranks the results based on the usefulness of the result to the user. It has the heuristics to order the search results based on the user need. HITS is a link analysis algorithm that rates Web pages [28]. The scheme assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. Topic-sensitive PageRank [29] is an extension of PageRank algorithm where a set of PageRank vectors are computed as opposed to only one to produce more accurate results. These vectors are related to the topics or the subject of the query rather than the web pages. They rank the web pages based on the relevance of the topic of the query contained in those web pages **[29].** *PopRank* is a domain-independent object-level link analysis model to rank the objects within a specific domain [30]. A popularity propagation factor is assigned to each type of object relationship effect of popularity propagation factors for these heterogeneous relationships over the popularity ranking. The two page ranking algorithms commonly used in Web structure mining are HITS and PageRank. In both algorithms all links from one node to another are considered equal when distributing rank scores. The weighted PageRank algorithm (WPR) is an extension to the standard PageRank algorithm. It considers both incoming links as well as the outgoing links of the pages. Both types of links are given equal importance while ranking. The rank scores are based on the popularity of the pages. WPR performs better than the conventional PageRank algorithm in terms of returning a larger number of

relevant pages to a given query [31]. A Semantic Web portal, called OntoKhoj [32] is designed to simplify the Ontology Engineering process. The methodology in developing OntoKhoj is based on algorithms used for searching, aggregating, ranking and classifying ontologies in Semantic Web. The proposed OntoKhoj would 1) allow agents and ontology engineers to retrieve trustworthy, authoritative knowledge, and 2) expedite the process of ontology engineering through extensive reuse of ontologies.

## 2.6 Bio2RDF: Linked Data for the Life Sciences

The Bio2RDF project is a network of the life science data which is coherently linked. Different databases across life sciences platform have been linked using open-source Semantic Web technologies to provide support biological knowledge discovery. It consists of different tools so that it can integrate the data from different databases which can then be queried using SPARQL which is an RDF Query Language [33]. It uses both syntactic and semantic data integration techniques, and presents the data on a distributed network server. It consists of 2 billion triples and is a publicly available system [34].

## 2.7 Linked Open Drug Data

In order to develop new drugs to cure diseases a large amount of biomedical data from various heterogeneous sources has to be integrated to get the relevant information about the disease to be treated. The new drugs to be developed are expected to be more effective against the disease than the existing drug and having less side effects. Linking Open Drug Data (LODD) project helps to bring the data sources together onto the Web of Linked Data and facilitates the integration of data. All the data and the datasets used from different sources have been strongly linked and also linked to the other Linked data. LODD contains 8.4 million RDF triples and the data from the data sources about drugs, Chinese

medicine, clinical trials, diseases and pharmaceutical companies have been linked together [35].

## 2.8 Drug Delivery and Polypharmacology

Drug discovery is a process of discovering and designing of drugs and is generally related to the fields of medicine, pharmacology and biotechnology. Drug discovery is a lengthy process and may take upto 15 years from initial stages of discovery to the time it is available for treating patients. The cost involved to develop a drug from scratch is as high as $800 million to $1 billion. Hence to develop an effective drug, it is necessary to understand how disease and infection are controlled at the molecular and physiological level. In order to do that the disease to be treated should be understood as well as possible, the underlying cause of the disease should also be found out. To understand the disease completely it is essential to understand as to which genes are affected by the disease as these genes in turn affect the proteins that they encode. Also since the proteins interact with each other in living cells, they affect the tissue in the areas in which the cells are located and as a result affect the patient on the whole. On gaining the complete understanding of the disease a target protein or gene is identified which is involved in the disease. Once the candidate for the new drug to be developed is identified, it is characterized and screened using high throughput screening etc. for the efficiency as a potential drug. Despite advances in technology and understanding of biological systems, drug discovery is still a lengthy, "expensive, difficult, and inefficient process" with low rate of new therapeutic discovery. Currently, the research and development cost of each new molecular entity (NME) is approximately US$1.8 billion. After considerable research, out of 5000 – 10000 compounds that are proposed for a potential drug for a disease only about 250 get approved

for preclinical trial, where they are tested in vivo as well as in vitro. This process takes about 3 - 6 years. Further out of the 250 compounds approved for pre-clinical trial only 5 get approved for the clinical trial which takes about 6 – 7 years and from that only one gets approved as a drug to cure the disease [1-4]. Figure 2.1 shows a typical drug discovery process. While designing a drug, natural products play a very significant role. The drugs could be plant derived, such as Belladonna, from the microbes such as streptomyces or marine invertebrates. Most drugs are derived from the natural products. However, despite the advances in the chemical synthesis techniques as well as combinatorial and cheminformatics, there has not been no increase in the number of drugs that are developed.

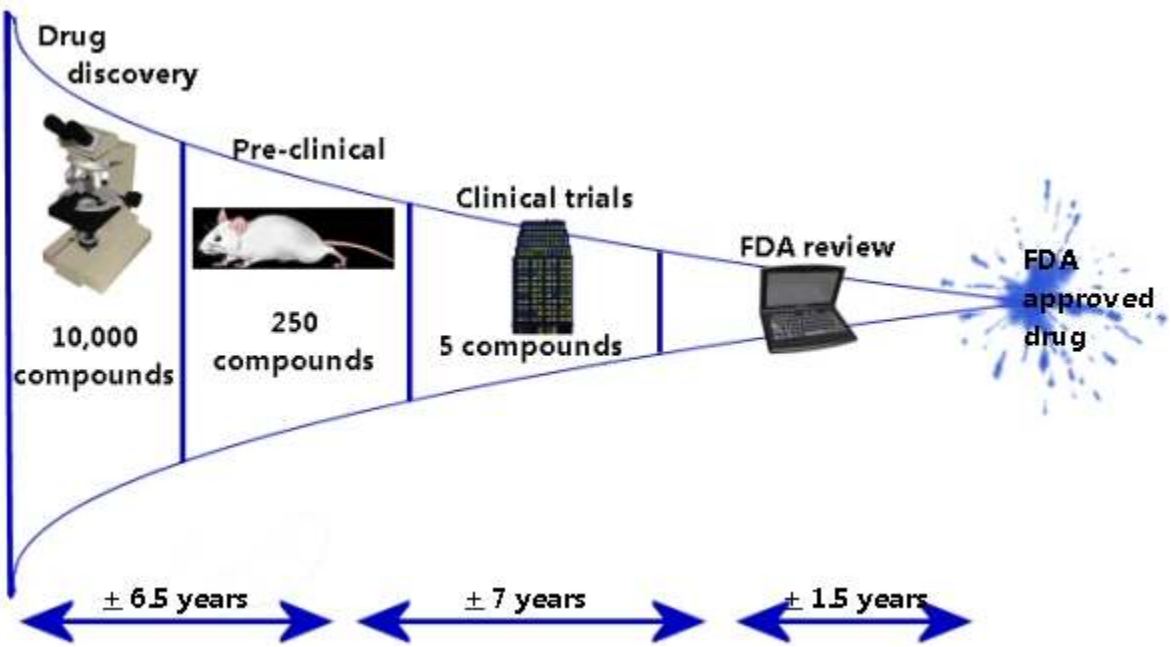


Figure 2.1. The Process of Drug Development [36]

## 2.9 Polypharmacology or Drug Repositioning

Polypharmacology is the study of interaction of drugs with multiple targets [37, 38]. Conventional drug design follows the principle of “one gene, one disease, one drug”. However, the same drug can be used with multiple diseases. Since the process of drug

discovery takes a long time of approximately 15 years and for every 5000-10000 proposed compounds, only one gets approved as a drug for the treatment of diseases and use on humans, the chances are that if a drug has been selected for a single target, it might be unsuccessful as it might be low in clinical efficacy. A drug can have a polypharmacological action on the targets that may fall in the same pathway. If the drug is able to interrupt the pathway at multiple points then it exhibits high efficacy. Also the already approved drugs may be less selective and may affect more than one protein target thereby affecting more than one disease. The examples of this drug repositioning are the cancer drug such as Gleevac (imatinib) [39] can bind to multiple kinases. Other examples are Propiomazine (Largon) and Promazine (Sparine) which have 14 targets each [40]. Also there might be certain diseases where a single gene or single protein may not be responsible for it. Hence, one drug may be insufficient to treat the disease. Thus the drug-target pair could be crucial for polypharmacology. There could be many-to-many relationships between the target – drug pairs. With the help of Semantic Web, the paths between the drug and the target can be found out and with the help of a suitable ranking technique, the most suitable drug for a target can be determined.

# CHAPTER 3. SYSTEM DESIGN

## 3.1 Overview

Large amount of data is available on the Internet and this data is ever growing. With the advancements in technology, there have been numerous data sources which are publicly available on the Internet. The biological and chemical fields have been of particular interest as they provide information about the humans. Numerous data sources are available for genes, proteins, genetic information, drugs, diseases etc. The study of the complex biological systems and the effect of drugs on these systems is possible by use of intelligent systems like Semantic Web. With the help of Semantic Web the available data can be integrated and meaningful results can be obtained [41,42]. The use of Semantic Web in life sciences, healthcare and drug discovery has already been demonstrated [43-45].

To study the relationships between the drugs and diseases and possibly to find the polypharmacological properties of drugs the data was taken directly from the websites and integrated to get the required data for the current study. The websites from where the data was taken are all open data sources. Many data sources are available on the Internet that contain similar and overlapping data about a particular entity. Hence, only one data source was chosen for each entity for the current study.

The Semantic Web has the advantage of integrating the relevant data from multiple, varied resources and possibly incompatible formats. The web resources are defined by Resource Description Framework (RDF) in the form of triples as <Subject, Predicate, Object>. The semantic associations are found by chaining of these triples. These associations are defined by ontologies. A complex semantic relation can be defined as the path between two entities where a path is defined as:

$$e_1, P_1, e_2, P_2, \ldots, e_{n-1}, P_{n-1}, e_n$$

where $e_i$ is an entity and $P_j$ is a predicate or property that defines a relationship between two entities, $e_i$ and $e_{i+1}$. Therefore this denotes a complex semantic relation between entities $e_1$ and $e_n$. Essentially this is the chaining of ontological triples, where the next triple has subject that was the object of the previous triple:

$<Subject_1, Predicate_1, Object_1> <Object_1, Predicate_2, Object_2>\ldots<Object_{n-1}, Predicate_n, Object_n>$

For Complex Relation discovery a search algorithm was the most important requirement which was robust and efficient as well as complete and optimal. The path result that is found by the search algorithm should be returned as a data type that was understandable by the program. The search algorithm should also be able to return multiple paths for comparison and ranking purposes. Finally the search algorithm needed to be constrained or bounded so that the results are meaningful. The Algorithm used in the present study used an Iterative Deepening Depth First Search to minimize space and time complexity and to provide complete and optimal complex paths. Complex relations are usually multi-hop relations, with goals probably being found at very low levels in the tree- therefore IDS returns deeper paths much faster- providing for more meaningful results.

Many association paths can exist between any two given entities. These paths are needed to be ranked to find the most relevant path within the system biology space. The entities can be associated with each other based on the context in which they are being considered. Semantic Web ranks different types of entities on the web. Different approaches have been employed to determine the ranking of each of these. Ranking of the

semantic relationships is different than ranking of documents. In Semantic Web the queries are ranked by finding the relevance in the keywords returned as the result of the query. The results of the query should be the association between entities which is the path between them. A number of different ranking algorithms have been reported in literature. PageRank [23] which analyses the incoming and outgoing links of documents, ObjectRank [24] an authority-based algorithm and ReConRank [26] a graph based ranking algorithm etc are few examples of ranking algorithms. In the present work, I propose to use the Iterative Deepening Depth First search Algorithm which runs repeatedly until it reaches the depth *d* where it finds the shallowest goal.

## 3.2 Data Sources

Numerous data sources are publicly available on the Internet for genes, genetic information, pathways, proteins, drugs and diseases. The information contained on many data sources may be similar and overlapping for the same entity. Besides this, the data may be present in various diverse formats such as text files, XML, different types of database formats etc. Integrating such systems could be a challenge which could be overcome with the help of Semantic Web. Integrating different systems could produce a network of linked drugs, proteins, genes and diseases. It should be possible to query such integrated systems to provide meaningful data. Efforts have been made in the past to integrate the biological and chemical data using Semantic Web. Some examples include LODD [46], YeastHub[47], LinkHub[48], BioDash[49] etc. For the current work the data on drugs was used from drugbank[50]. It is the most widely used database on drugs. It has the links to other databases such as UniProt[51], GeneCards[52] and OMIM[53] which were also considered for study.

Protein data was obtained from UniProt[51] which is a widely used database for proteins and protein sequences. It has been used by National Center of Biotechnology Information (NCBI) of National Institute of Health (NIH). The proteins which acted upon by the drugs in drugbank are identified by the UniProt ID which links them to the protein database. Gene data was obtained from GeneCards[52] which contains information about all known and predicted human genes and information on diseases was linked in OMIM database[53].

### 3.2.1 Integration of Data Sources

It is very important to define the relationships between the entities accurately as well as comprehensively. This is because the entities are taken from the sources where their relationships may be very poorly and loosely defined. The ontology was developed by integrating the data available online from DrugBank, UniProt, GeneCards and the OMIM websites. This was done in order to establish the connectivity or the relationship between the drugs, diseases, protein targets and the genes. Also called as Ontology mapping, integration of data sources is an important requirement of this work as the data has been used from different websites.

## 3.3 Methods of Complex Relation Discovery

### 3.3.1 Algorithm

A number of search algorithms have been reported in literature. The present work uses the Iterative Deepening Depth First Search Algorithm. It searches iteratively increasing the depth every time until the goal is reached. It is a state space search. It has the advantage of both the Breadth First Search as well as the Depth first search as it combines the state space efficiency of Depth First Search and completeness of the Breadth First

Search. The space complexity of IDDS is O(*bd*) where *b* is the branching factor and *d* is the depth of the shallowest goal. The time complexity of IDDS is the same as Depth First Search O($b^d$). In this type of search, the nodes on the lowest level are expanded once, those on the next to lowest level are expanded twice, and so on, up to the highest level or root of the search tree, which is expanded *d* + 1 times[54]. So the total number of expansions in an iterative deepening depth first search is given as

$$(d+1)b^0 + (d)b^1 + (d-1)b^2 + \ldots 2b^{d-1} + 1b^d$$

This algorithm was used to minimize space and time complexity for a finite graph and to provide complete and optimal complex paths. Complex relations can be represented as a graph with a start node and a goal node. The Iterative Deepening Depth First Search algorithm can be implemented as follows

```
IDDFS(start, goal)
{
  depth = 0
  while(!true)
  {
     result = search(start, goal, depth)
    depth = depth + 1
  }
  return result
}

search(start, goal, depth, visited)
{
  if ( depth >= 0 )
   {
    if (start == goal )
      return start

        visited.insert(start)


    for each child in expand(node)
     if(!visited)
      search(child, goal, depth-1, visited)

  }}
```

Algorithm 1. Iterative Deepening Algorithm for Relation Discovery [55]

Different methods can be used to find a complex relation between two entities. An important measure for these complex relations is the number of hops it takes to reach $e_n$ from $e_1$. For the current work the number of hops is limited to 4,

## 3.4 Problem Statement

Drug development has been a challenging field for a long time. It presents the problem of high cost, long development time and low success rate. With the advancements in technology, numerous data sources are available for genes, proteins, genetic information, drugs, diseases etc. These can be used to study the polypharmacological properties of the drugs to reduce the cost of drug development. The available data is disparate in nature. Integrating the heterogeneous data to find meaningful results has always been a challenge. To achieve this, heterogeneous data across different knowledge domains needs to be linked in such as way as to effectively find the associations of the drugs to potential new uses.

## 3.5 Contribution to the Paper

In this work an attempt is made to integrate the heterogeneous data over two different domains – chemical and biological to find the relationship between the drugs and diseases. The use of Semantic Web to integrate the heterogeneous data over these domains is demonstrated. For this purpose the search algorithm was implemented using a simple programming language and the associations between the drugs and diseases can be retrieved using SQL rather than complex methods.

## 3.6 Experimental

### 3.6.1 Omim Model

This section discusses the implementation of software to discover complex semantic relations: The main role of Omim is to find complex relationships between two input entities namely a drug and a disease.

### 3.6.2 Software Requirements

Python 2.7.2: The system developed in this study involves a relational database which was created by downloading the flat files from various websites. The data was parsed using Python. Python is general purpose, high level programming language. It has a fully dynamic type system and automatic memory management. It is used as a scripting language for web applications. Python 2.7.2 was used for the current study [56-58].

Django: Django is a high-level Python Web framework that helps in rapid application development and clean design [59]. It consists of object relational mapper to help define the data models entirely in Python giving a rich dynamic database API. It provides an automatic design interface to help add and modify contents. Django 1.3 version was used for the development of the application [59].

SQLite: SQLite is a software library the source code of which is in public domain. It is the most widely used serverless self-contained SQL database engine [60]. It is compatible with most of the programming languages such as C, C++, C#, Visual Basic, Java, Python etc. The version of SQLite used for the application was 3.7.8.

The system involves a relational database which was created by downloading the flat files from various websites and only the data needed for the study was used out of the

entire flat file. The data required for the study was the drug name, the DrugID, the UniProtID and the information of the Proteins, the GeneID as well as the MIM# for the disease and the disease name. Figure 3.1 shows the Omim Architecture.

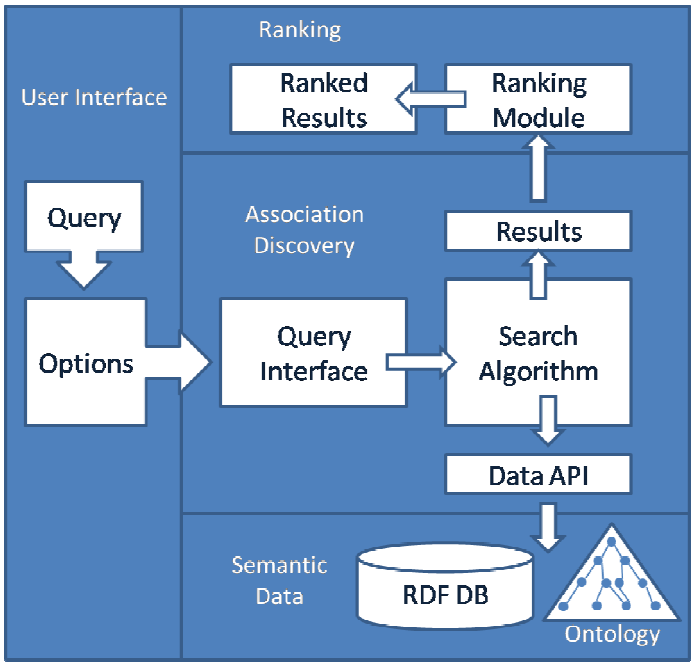


Figure 3.1. The Omim Component Architecture

3.6.3 Data Sources

Most of the data used for this work was taken from DrugBank website [8,61,62] which is a freely available public resource. The DrugBank database is a bioinformatics and cheminformatics resource. It combines the chemical, pharmacological and pharmaceutical information of a drug with comprehensive drug target information including protein and gene sequences, structure, and pathway information. There are 6707 drugs listed in DrugBank. The information about each drug is given in a DrugCard which contains more than 150 data fields with information of the drug/chemical data as well as the drug target or protein data. Of these fields only the Generic Name of the drug and the Accession number

which was used as an Identifier of the drug were used. Some of the data taken directly from DrugBank is presented below in Table 3.1.

| **Identification** | | |
|---|---|---|
| Name | **Tacrine** | |
| Accession Number | **DB00382** (APRD00690) | |
| Weight | Average: 198.2637 | |
| Chemical Formula | $C_{13}H_{14}N_2$ | |
| **Properties** | | |
| State | solid | |
| Melting point | 183.5 $^{o}$C | |
| External Links | **Resource** | **Link** |
| | KEGG Compound | C01453 |
| | PubChem Compound | 1935 |
| | Drugs.com | http://en.wikipedia.org/wiki/Tacrine |
| | Wikipedia | http://en.wikipedia.org/wiki/Tacrine |
| **Targets** | | |
| **1. Acetylcholinesterase**<br>Pharmacological action: **yes**<br>Actions: **inhibitor**<br>Rapidly hydrolyzes choline released into the synapse<br>Organism class: **human**<br>UniProt ID: P22303<br>Gene: ACHE<br><br>**2. Cholinesterase**<br>Pharmacological action: **yes**<br>Actions: **inhibitor**<br>An acylcholine + H(2)O = choline + a carboxylate<br>Organism class: **human**<br>UniProt ID: P06276<br>Gene: BCHE | | |
| **Enzymes** | | |
| **1. Cytochrome P450 1A2**<br>**Actions: substrate**<br>**UniProt ID: P05177** | | |

Table 3.1. Sample Data from DrugBank

The information about the Protein Targets and the Enzymes was taken from UniProt website [51]. UniProt [63-65] is a freely accessible knowledgebase that provides high quality protein sequence and functional information to the scientific community mostly dealing with the biological data. Out of the different databases in UniProt only

UniProtKB [65] was used for this study. In release 2010_09 of 10 August 2010, UniProtKB/Swiss-Prot contained 519,348 entries [65]. Taking the above example from the drugbank it is evident that one of the Protein Targets is Acetylcholinesterase with UniProt ID as P22303 and Gene as ACHE. For the above UniProtID the data for Protein Target is given. The data given below in Table 3.2 is a part of data that is present in the UniProt website against UniProtID P22303. It is also found that this ID has the reference to Tacrine from DrugBank**.**

| **P22303** (ACES_HUMAN) Reviewed, UniProtKB/Swiss-Prot | | |
|---|---|---|
| Protein names | *Recommended name:* | |
| | Acetylcholinesterase | |
| | Short name=AChE | |
| Gene names | Name: | **ACHE** |
| Organism | **Homo sapiens (Human)** | |
| **Protein attributes** | | |
| Sequence length | 614 AA. | |
| **Cross-references** | | |
| **Sequence databases** | | |
| EMBL | M55040 mRNA. Translation: AAA68151.1. | |
| GenBank | S71129 Genomic DNA. Translation: AAC60618.1. Sequence problems.. | |
| **3D structure databases** | | |
| **Protein-protein interaction databases** | | |
| IntAct | P22303. 6 interactions. | |
| **Genome annotation databases** | | |
| GeneID | 43 | |
| KEGG | hsa:43. | |
| UCSC | uc003uxd.1. human. | |
| **Organism-specific databases** | | |
| CTD | 43 | |
| GeneCards | GC07M095117. | |
| MIM | 100740. gene+phenotype. 112100. phenotype. | |
| **Other** | | |
| **Drugbank** | DB00989. Rivastigmine. | |
| | DB00382. Tacrine. | |

Table 3.2. Sample Data from www.uniprot.org

GeneCards is a database of human genes that provides genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes [66-68]. Information that is present in GeneCards includes disease relationships, mutations, gene expression, gene function, pathways, protein-protein interactions, related drugs & compounds as well as links to other information such as antibodies, recombinant proteins, clones etc. GeneCards was used to find the details of the gene involved in the disease. The total number of GeneCards entries is 67217. Taking the same example of Tacrine above, the gene is given as ACHE which has the cross-reference to GeneCards website. Table 3.3 gives the sample data from GeneCards website

<table>
<tr><td>ACHE Gene</td><td colspan="5">acetylcholinesterase<br>(Previous names: acetylcholinesterase (YT blood group), acetylcholinesterase (Yt blood group) )</td></tr>
<tr><td>GC07M100487</td><td colspan="5">(Previous symbol: YT)</td></tr>
<tr><td rowspan="4">Aliases & Descriptions for ACHE gene<br><br>(According to HGNC, Entrez Gene, UniProtKB/Swiss-Prot, OMIM,</td><td colspan="5">Aliases & Descriptions</td></tr>
<tr><td colspan="4">acetylcholinesterase</td><td>OTTHUMP00000211347</td></tr>
<tr><td>External Ids:</td><td>HGNC: 108</td><td>Entrez Gene: 43</td><td>Ensembl: ENSG00000870857</td><td>UniProtKB: P22303</td></tr>
<tr><td colspan="5">Export aliases for ACHE gene to outside databases</td></tr>
<tr><td colspan="5">Previous GC identifers: GC07M099022 GC07M100085 GC07M100099 GC07M100132 GC07M100325 GC07M095117</td></tr>
<tr><td>Summaries for ACHE gene<br>(According to Entrez Gene, Wikipedia's Gene Wiki, UniProtKB/ Swiss-Prot</td><td colspan="5">Entrez Gene summary for ACHE:<br>physostigmine used to treat glaucoma. AChE inhibitors are also used in the management of mild to moderate Alzheimer's disease.<br>Gene Wiki entry for ACHE (Acetylcholinesterase)</td></tr>
<tr><td>Gene Function for ACHE gene<br>(According to UniProtKB, Animal models from MGI May 11 2011,</td><td colspan="5">Function Summary:<br>UniProtKB/Swiss-Prot: ACES_HUMAN, P22303<br>Function: Terminates signal transduction at the neuromuscular junction by rapid hydrolysis of the acetylcholine</td></tr>
</table>

Table 3.3. Sample Data from www.genecards.org

The drug obtained from the drugbank website was linked to the disease through a path. The details of the disease to which it is linked is found in OMIM**.** Online Mendelian Inheritance in Man (OMIM) is a database that consists of all the known diseases with a genetic component. It also links these diseases to the relevant genes in the human genome. All diseases may not be linked to the human genome as it is still in the developmental stage. OMIM provides references for further research and tools for genomic analysis of a catalogued gene [12,69]. It is one of the databases that have been used by U.S. National Center for Biotechnology Information (NCBI). The Protein target Acetylcholinesterase was linked to Alzheimer's disease as Tacrine is a drug that is used for the treatment of Alzheimer's disease at nearly all stages.

For the current study the following schema given in Figure 3.2 was used to link the entities from various sources: The drug name, properties, enzymes and protein targets obtained from DrugBank are linked with the genes to find out which genes are affected by the disease.

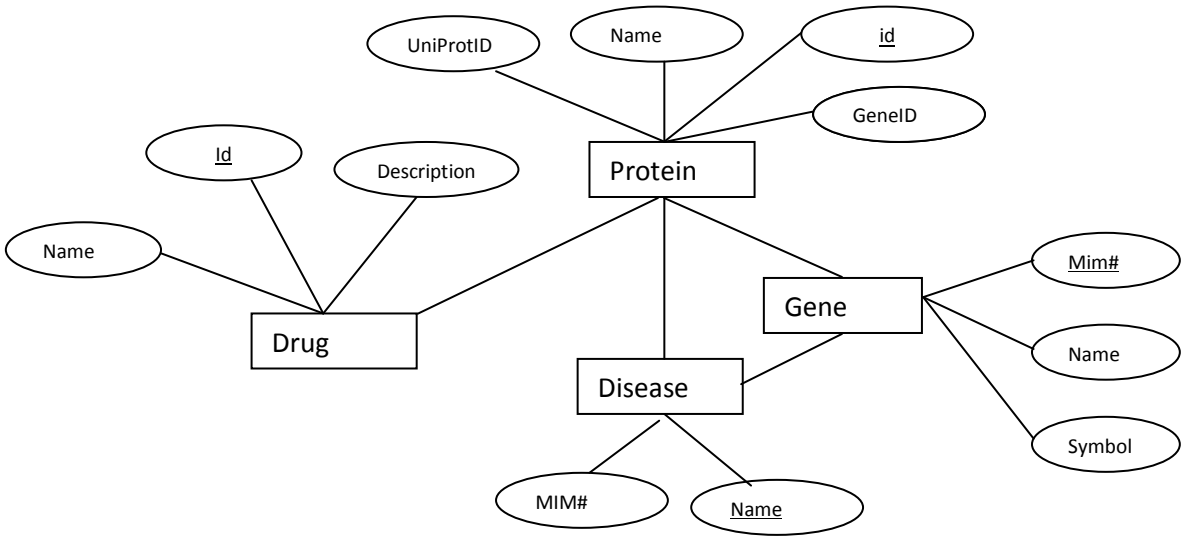


Figure 3.2. Schema

### 3.6.4 Ontological Management

Ontology of the drugs and the diseases as well as the intermediaries was developed. Protégé was used to develop the ontology. Ontology mapping was carried out to find the relation between the drugs and the diseases as following:

1. The Drug was taken from DrugBank website [50]. The drug had a unique Identifier the Drugbank ID by which it is identified. The website contains all the information about the given drug.
2. The drug has mechanism of action on the disease which is carried out with the help of target proteins and enzymes. The drug may inhibit the action of a protein.
3. The protein targets and the enzymes have UniProtID. They are related to genes with the help of GeneID. Each protein target and enzyme is transcribed by a gene.
4. Each protein target and enzyme also contains a unique MIM#. With the help of this MIM# it is related to the disease it affects. The information of disease is given in the OMIM website. However, all genes on the human genome have not been mapped to OMIM. Hence, the disease name was used to link to the gene and drugs.

To study the relationship between drugs and diseases a high level ontology suitable for the current work was developed. The goal here was to devise a drug- and disease-centric knowledge framework that is helpful for both data integration and knowledge discovery. Ontology of the drugs and the diseases as well as the intermediaries was developed. Only the fields that were considered for study were taken to develop the ontology. The other fields could be obtained from the respective website just like in a database management system. The required number of classes were defined for this work which could give the semantic relationship between the drug and the disease including the

biological entities that a disease may affect such as the genes. Hence 6 different classes with 6 properties were defined. The ontology developed here consists of different triples for every drug to disease relationship as given in Figure 3.3. With the vast data available on the Internet and with the growing nature of the data, no ontology can be complete. However, the work needs to be simplified so that the integration of the existing and future data could be simplified and the desired semantic associations could be obtained.

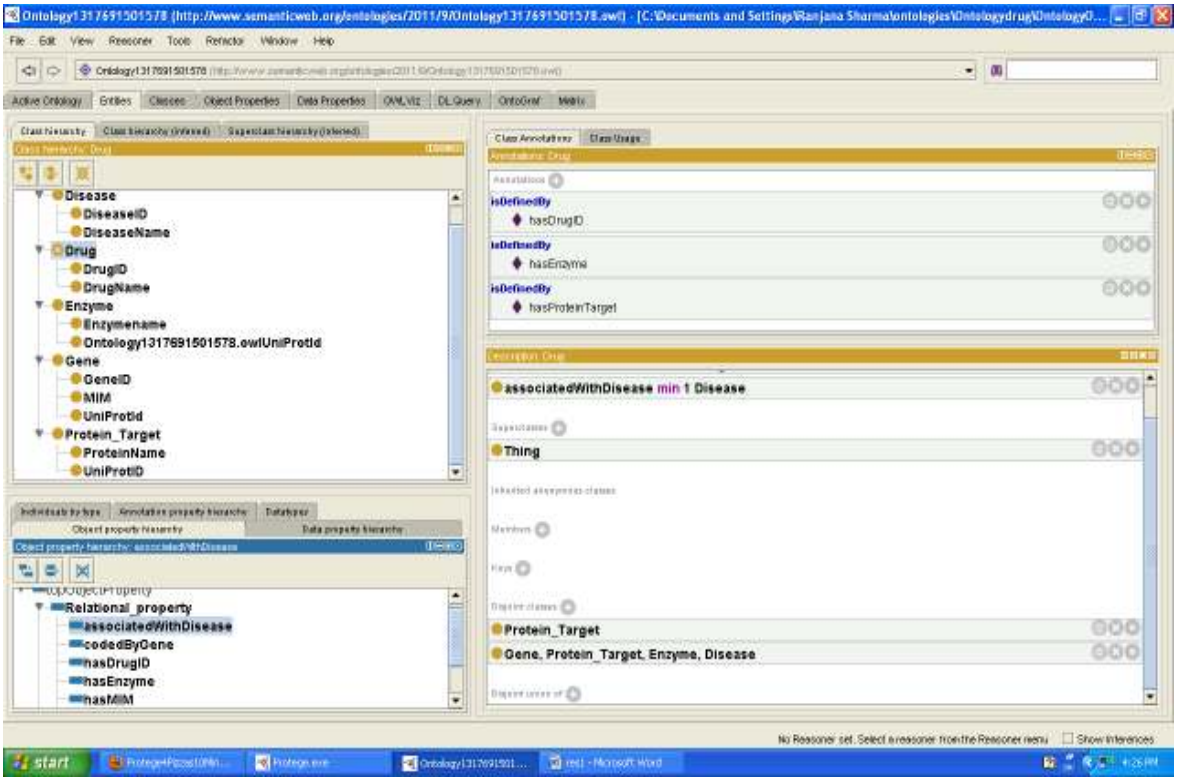

Figure 3.3. Screenshot of the Ontology from Protégé

Ontology mapping or integration of data sources was carried out by identifying the drug from DrugBank by accession number; identifying the protein that it inhibits; associating the drug and drug targets to the genes and subsequently with the diseases.

The ontological organization is handled by the relational database and manages entities, predicates, and triples made up of those entities and predicates. All entities and

predicates are identified by a unique key that is taken from the different websites. This key is indexed in the database to provide performance for the searching. The user manages the ontology from the interface through individual components for entities, predicates, and triples so that the management happens separately from all other components.

The drug related data compiled from DrugBank contains comprehensive information about Protein targets and enzymes related to drugs. The data set contains 6707 drugs out of which 1436 are FDA approved. The present system parsed the data of all drugs for the integration with other data. The protein data was also parsed and was mapped to the gene data from GeneCards as well as OMIM websites for mapping to different diseases. The application found the associations between drug and protein targets as well as enzymes. The number of protein targets/enzymes/transporter/carrier of the drugs in drug bank was found to be 4228. Each protein and enzyme is coded by a gene. The details of the proteins can be obtained from the UniProt database using the UniProtID.

The Algorithm used here is Iterative Deepening Depth First Search which searches all its neighbors until the goal is reached. This algorithm was used to minimize space and time complexity for a finite graph and to provide complete and optimal complex paths. Complex relations can be represented as a graph with a start node and a goal node. Different methods can be used to find a complex relation between two entities. An important measure for these complex relations is the number of hops it takes to reach $e_n$ from $e_1$. However, since this work is limited to finite set and the number of hops are limited to a maximum of 4.

**Example1: Complex relation discovery for antimalarial drugs**

The main goal of this paper is to discover complex semantic relationships between drugs and diseases as well as the target proteins or enzymes and the genes that are affected by those proteins or enzymes. If the relationship exists through intermediate entities it can be expressed as a chained triple. Every element in a RDF triple is a Uniform Resource Identifier (URI). When two resources have the same URI they are said to be identical and all data for identical resources is merged.

Various antimalarial drugs that are commonly used for the treatment of malaria were used to carry out the study. A list of the antimalarial drugs is compiled from the literature [70] and used in study. Some drugs have been indicated as having more than one use and malaria being the new indication for such drugs were also considered for study. The Protein Targets, GeneID and MIM# were taken from drugbank. Thus Chloroquine commonly used for treatment of malaria was taken from drugbank, has the accession number as DB00608. The drug has four targets as Ferriprotoporphyrin IX, Glutathione S-transferase A2, Tumor necrosis factor, Toll-like receptor 9. The UniProt ID is given gives the detailed information about the protein and the GeneID gives the detailed information about the gene from the GeneCards website. The relationship can be defined as chained triples. The Table 3.4 below shows the relationship between the drug, Chloroquine, and the disease, Malaria, using MIM#.

| Subject | Predicate | Object |
|---|---|---|
| Chloroquine(DB00608) | hasProteinTarget | Tumor Necrosis Factor |
| Tumor Necrosis Factor | hasUniProtID,<br>hasAssociatedGene | P01375,<br>TNF |
| TNF | hasMIM# | 191160 |
| 191160 | associatedWith | Malaria |

Table 3.4. Relationship between the Drug and Disease through MIM#

However, in the example below in Table 3.5, the gene, DHFR, treats the disease by acting on the metabolism of the parasite, Plasmodium falciparum and Plasmodium vivax, known to cause malaria. Hence, this gene h**as** been mapped to OMIM using the MIM# that corresponds to the metabolism rather than directly to Malaria. Since, the disease name was used, the application found a relationship directly with the disease

| Subject | Predicate | Object |
|---|---|---|
| Proguanil(DB01131) | hasProteinTarget | Dihydrofolate reductase |
| Dihydrofolate reductase | hasUniProtID,<br>hasAssociatedGene | P00374,<br>DHFR |
| DHFR | associatedWith | Malaria |

Table 3.5. Relationship between Drug and Disease without the Use of MIM#

Some of the known antimalarial drugs contain Ferriprotoporphyrin IX. This compound is known to form cytotoxic complexes with the antimalarial drugs that cause plasmodial membrane damage. However, this compound does not have any UniProtID, GeneID associated **wit**h it and there is little information about this compound on drugbank. But this compound has been known to be used in the treatment of malaria. Table 3.6 shows direct relationship with the disease.

| Subject | Predicate | Object |
|---|---|---|
| Chloroquine (DB00608)<br>Halofantrine (DB01218)<br>Mefloquine (DB00358)<br>Primaquine(DB01087)<br>Amodiaquine (DB00613)<br>Quinine (DB00468) | hasProteinTarget | Ferriprotoporphyrin IX |
| Ferriprotoporphyrin IX | associatedWith | Malaria |

Table 3.6. Relationship between Drug and Disease Directly through Protein

**Example 2. Finding relations to demonstrate Polypharmacology of drugs**

The drugs that are already effective against one disease can also be used for the treatment of other diseases. Using the same drug against other diseases can be highly beneficial [37-39]. Some drugs have been reported from the literature [71,72] with original and new uses which are different than their original uses for which they had been approved. Thalidomide had been earlier approved as a drug for sedation, nausea and insomnia. It is also being used in the treatment of multiple myloma [73]. Acetylsalicylic acid, commonly called as Aspirin [74,75] has been used as analgesic and antipyretic to reduce aches and pains and fever for a long time. It is also being used in cardiology to prevent heart attacks, strokes and blood clotting due to its antiplatelet activity [76]. Also Miltefosine used in the treatment of Cancer [77] has the new indication as Visceral leishmania [78]. A very common example is of Sildenafil which was earlier used for hypertension [79] is now being used for male erectile dysfunction in the name of Viagra [80].

Polypharmacological properties of the drugs can be studied by finding the relation of drugs to proteins and genes and the diseases. The relationships may involve many different paths originating from one disease to a drug and vice-versa with intermediate genes and protein targets. If two drugs have at least two same targets, they will show the

polypharmacological properties. Alzheimer's disease is generally treated by inhibiting the enzymes Acetylcholinesterase and Cholinesterase. The gene that codes Acetylcholinesterase is ACHE which is a known factor in Alzheimer's disease and the gene that codes Cholinesterase is BCHE**.** There are a number of approved drugs in drugbank that are known to act as the inhibitors of Acetylcholinesterase as well as Cholinesterase. Of the 16 approved drugs in drugbank 4 drugs, namely, Tacrine, Rivastigmine, Galantamine and Choline contained both Acetylcholinesterase as well as Cholinesterase and one Donepezil contained Acetylcholinesterase. Tacrine is given as an example in Figure 3.4.

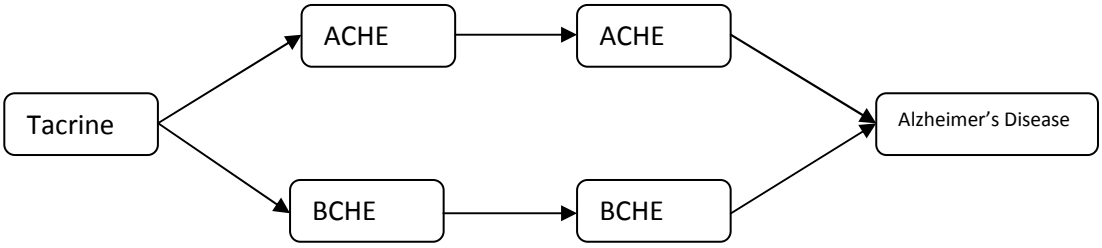


Figure 3.4. More than 1 Path from Drug to Disease

The experimental drugs such as Methylphosphinic Acid, 2-(N-Morpholino)-Ethanesulfonic Acid or Fucose could be the potential novel drugs and may show the polypharmacological properties. Methylphosphinic Acid has both Acetylcholinesterase as well as Cholinesterase as the principal components. Hence, it could be used for the treatment of Alzheimer's Disease whereas the other two act on as many as 30 proteins. They should be studied for more than one indication as they are being developed. They might be helpful in treating more than one disease including Alzheimer's Disease, Myasthenia Gravis and Glaucoma which also involve Acetylcholinesterase.

**Example 3. Finding Drug-gene-disease relationship**

Since at least one gene is involved in all the diseases, an attempt was made to find the relationship between drug, gene and the disease. The drug-gene-disease relationship was studied for the gene ACHE. Alzheimer's disease is generally treated by inhibiting the enzyme Acetylcholinesterase. The gene that codes Acetylcholinesterase is ACHE which **is** a known factor in Alzheimer's disease**.** There are a number of approved drugs in drugbank that are known to act as the inhibitors of Acetylcholinesterase. These drugs could be studied for their potential use for the treatment of Alzheimer's disease. A study is carried out using a gene ACHE, which had been involved in a number of diseases. It was found that as many as 50 drugs act on ACHE some of them still being in the experimental stage. Only approved drugs were considered for study. Although all the drugs act as the inhibitors of Acetylcholinesterase there are some drugs that help in the treatment of Alzheimer's disease, some for Myasthenia Gravis and some for Glaucoma (Figure 3.5).

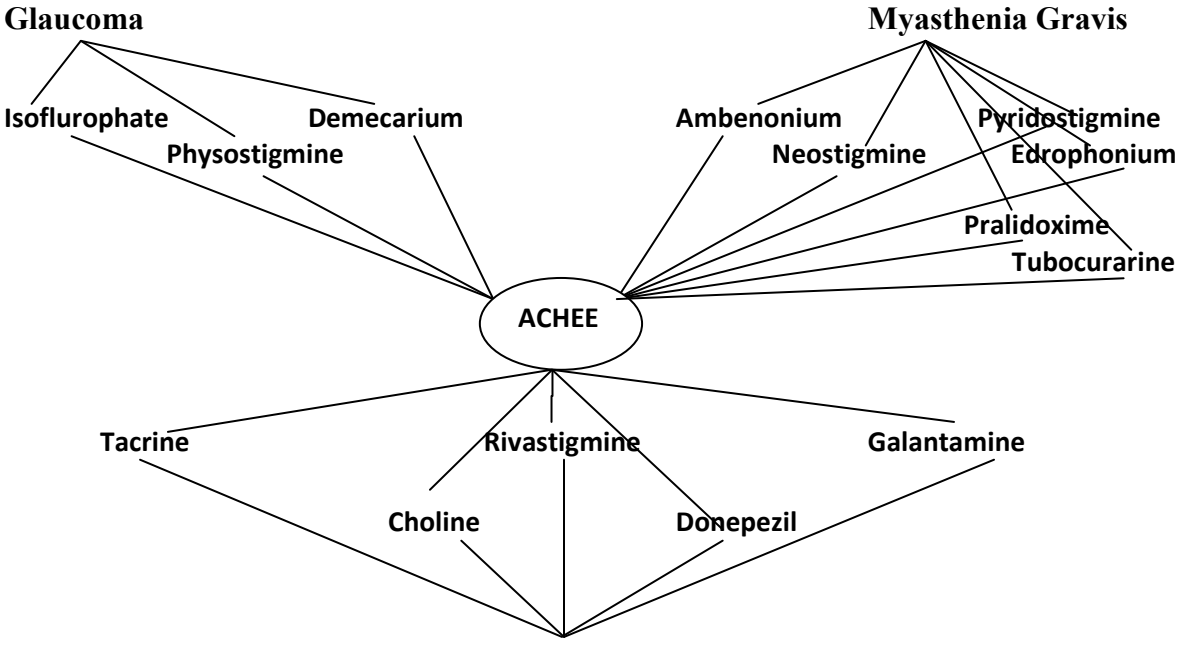


Figure 3.5. The Drug-Disease-Gene Relationship for Gene ACHE

Protein is a chain of polypeptides linked together. A polypeptide is a linear chain of amino acids linked together by means of a peptide (- H – N – CO -) bond. Enzymes are proteins that catalyze a chemical reaction in the human body. A polypeptide chain in a protein is coded by a particular gene. A gene is made up of a genetic code. The sequence of a genetic code results in the sequence of amino acids in a polypeptide. In other words each gene codes a polypeptide which is contained in a protein. The same polypeptide may be contained in more than one protein. Different polypeptides contained in one protein may be synthesized by different genes. Hence, a many-to-many relationship exists between proteins and genes where as a one-to-one relationship exists between a polypeptide and a gene. A gene may be involved in causing a disease by synthesizing a protein/enzyme. A drug acts as an inhibitor for the action of a protein to cure a disease. It is assumed that if a gene has been proved to cause a disease and if a drug cures that disease, then if the gene is involved in another disease, then the same drug may find potential use in the treatment of that disease. This needs further study. In the current study, three drugs namely Tacrine that has been used primarily for the treatment of Alzheimer's disease, Pyridostigmine that has been used for the treatment of Myasthenia Gravis and Demecarium used for the treatment of Glaucoma were considered. As per assumption any of the above drugs could be used for the treatment of the other disease as well. However, this assumption needs further study.

# CHAPTER 4. CONCLUSION AND FUTURE WORK

## 4.1 Conclusion

A large number of diseases have been known to affect humans every day. Computational methods have been proposed to find the disease genes, protein interaction etc. However, there have been few methods that have been proposed to facilitate drug development. Large amount of biological and chemical data is present on the Internet in public databases. As the Internet expands to contain more and more data, information definition and searching is going to become increasingly important. Soon it will be impossible for humans to digest all the content on the web by themselves.

To solve the aforementioned problem, advantage of the recent Semantic Web technologies was taken to solve drug discovery problem. By extracting metadata of web resources, the current system can now perform more complex searches and discover more complex relationships between data that only human users could once do. An intelligent semantics-based searching scheme has been proposed in this paper that discovers complex semantic relations between the drugs and diseases. In this proposed scheme, complex semantic relations, identified by the chaining of the ontological triples in the metadata will allow to identify links in seemingly unrelated resources or ontological structures. The Iterative Deepening Depth First Search algorithm along has been used along with ranking and other heuristic pruning algorithms. The system described in this paper has the ability to parse the data from the websites and only the data necessary for the current work is used. This makes handling of the enormous data on the Internet easier. Real web data has been used to test the proposed system.

## 4.2 Future Work

As drug discovery is a very expensive and time-consuming process, polypharmacology is a better option for the treatment of the diseases. The effect of the existing drugs on the biological systems could be studied for their new indications and their potential use. The biological and chemical data present on the Internet can be studied and in combination of systems biology, chemogenomics and bioinformatics the drugs with the new indications could be developed. The current system can be easily extended to use an efficient data management system in order to maintain a high performance.

1. In the present work, the data is used directly from the websites. This work can extended by incorporating more websites that contain other comprehensive data about the diseases such as their indications, the phenotype information, the pathways they affect, the chemical structure of the drugs etc and finding the semantic relations between them. The diseases as well as drugs could be studied for their similarities and clustered together. The polypharmacological properties of the drugs affecting one disease could be studied so that another disease with the similar indications may have the possibility of being treated with existing drugs.

2. The search algorithm can be updated and a more optimized search algorithm is needed which could include weights in the algorithm and also studies could be carried out in a particular context.

3. The ranking algorithm can be implemented to find more appropriate drug for a particular disease. The polypharmacological property of the drug could be studied

in a particular context. The ranking algorithm could include the context as well as the other heuristics.

All this is possible because of the web technology used and because of the component nature of the software- plus the planned requirements changes were included in the code design and architecture.

REFERENCES


[1] DiMasi, J.A., Hansen, R.W., & Grabowski, H.G., "The Price of Innovation: New Estimates of Drug Development Costs," *Journal of Health Economics*, 22, 151-185, 2003.

[2] DiMasi, J.A., Hansen, R.W., Grabowski, H. G., Lasagna, L., "Cost of innovation in the pharmaceutical industry", *Journal of Health Economics,* 10, 107-142, 1991.

[3] Masia, N., "Cost of Developing a new Drug". , US Dept of state publication – Focus on Intellectual Property rights, <http://www.america.gov/st/econ-english/2008/April/20080429230904myleen 0.5233981.html> , 2008.

[4] Collier, R., "Drug development cost estimates hard to swallow", *CMAJ*, 180(3), 279–280, 2009.

[5] Sleigh, S.H., Barton, C.L., "Repurposing Strategies for Therapeutics", *Pharm Med,* 24 (3), 151-159, 2010.

[6] Aronson, J. K. "Old drugs – new uses", *Br J Clin Pharmacol*. 64(5), 563–565**,** 2007**.**

[7] Ashburn, T. T. and Thor, K. B., "Drug repositioning: Identifying and developing new uses for existing drugs", *Nature Review: Drug Discovery*, 3, 673-683, 2004.

[8] Knox, C., Law,V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C. & Wishart, D. S. "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs", *Nucleic Acids Research*, *Database issue*, 39, D1035–D1041, 2011,

[9] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M., "KEGG for representation and analysis of molecular networks involving diseases and drugs", *Nucleic Acids Res*., 38, D355-D360, 2010.

[10] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., "From genomics to chemical genomics: new developments in KEGG", *Nucleic Acids Res*. 34, D354-357, 2006.

[11] Kanehisa, M. & Goto, S., "Kyoto Encyclopedia of Genes and Genomes", *Nucleic Acids Res*. 28, 27-30, 2000.

[12] Hamosh, A., Scott, A., Amberger, J., Bocchini, C., McKusick, V. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders", *Nucleic Acids Research: Database issue"*, 33, D514–D517, 2004.

[13] Berners-Lee, T., Hendler, J. and Lassila, O., "The Semantic Web", *Scientific American Magazine*, May 2001

[14] Candan, K.S., Liu, H. and Suvarna, R., "Resource Description Framework: Metadata and Its Applications", *ACM SIGKDD Explorations Newsletter*, 3(1), 6-19, 2001.

[15] "Resource Description Framework (RDF): Concepts and Abstract Syntax," ed. Klyne, G., Carroll, J. J., W3C, <.http://www.w3.org/TR/2004/ REC-rdf-concepts-20040210/ , 10 February 2004,

[16] Anyanwu, K., & Sheth, K., "The ρ Operator: Discovering and Ranking Associations on the Semantic Web ", *ACM SIGMOD Record : Special issue on Amicalola Workshop*, 31, n4, 2002.

[17] Kemafor, A., & Sheth, A., "ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web", *Proceedings of 12th International Conference on World Wide Web,* Budapest, Hungary, 680-683, 2003.

[18] Aleman-Meza, B., Halaschek, C., Arpinar, I. B., & Sheth, A. "Context-Aware Semantic Association Ranking", *Proceedings of Semantic Web and Databases Workshop-03,* Berlin, 33-50, 2003.

[19] Aleman-Meza, B., Halaschek-Wiener,C., Arpinar, B., Ramakrishnan, C., & Sheth, A., "Ranking Complex Relationships on the Semantic Web ", *IEEE Internet Computing* , 9 (3), 2005.

[20] Halaschek, C., Aleman-Meza, B., Arpinar, B., & Sheth, A. P." Discovering and Ranking Semantic Asscociations over a Large RDF Metabase", *Proceedings of the 30th VLDB Conference*, Toronto, Canada, 1317-1320, 2004.

[21] Altman, R.B., "Building successful biological databases", *Brief. Bioinformatics,* 5 (1): 4–5, 2004.

[22] Khatri, P., Done, B., Rao, A., Done, A. & Draghici,S., "A semantic analysis of the annotations of the human genome", *Bioinformatics*, 21(16 ) 3416–3421, 2005.

[23] Page, L., Brin, S., Motwani, R., & Winograd, T., "The PageRank citation ranking: Bringing order to the Web", *Stanford Digital Libraries Working Paper,* 1998.

[24] Balmin, A., Hristidis, V., Papakonstantinou, Y., "ObjectRank: Authority-Based Keyword Search in Databases", *Proceedings of the Thirtieth International Conference on Very Large Data Bases",* 30, 564-575, 2004.

[25] Davison , B.D., Gerasoulis, A. , Kleisouris , K., Lu ,Y., Seo ,H.J., Wang ,W., Wu, B., "DiscoWeb: Applying link analysis to web search", *Poster Proceedings of the Eighth International World Wide Web Conference,* 1999.

[26] Hogan, A. ,Harth,A., Decker, S., "ReConRank: A Scalable Ranking Method for Semantic Web Data with Context", *2nd Workshop on Scalable Semantic Web Knowledge Base Systems,* 2006

[27] Anyanwu, K., Maduko, A. & Sheth,A., "SemRank: Ranking Complex Relationship Search Results on the Semantic Web", *Proceedings of the 14th International World Wide Web Conference, Chiba, Japan*, 2005.

[28] Kleinberg, J. "Authorative sources in a hyperlinked environment", *J. ACM*, 48,604-632, 1999.

[29] Haveliwala, T. H., "Topic-Sensitive PageRank", *Eleventh International World Wide Web Conference*, Honolulu, Hawaii, 2002.

[30] Nie, Z., Zhang,Y., Wen,J.R., Ma, W.Y., "ObjectLevel Ranking: Bringing Order to Web Objects", *Proceedings of 14th International World Wide Web Conference, Chiba, Japan,* 2005.

[31] Xing, W., Ghorbani, A., "Weighted PageRank algorithm", *Proceedings. Second Annual Conference on Communication Networks and Services Research,.* 305 – 314, 2004.

[32] Patel, C., Supekar,K., Lee,Y., Park,E.K., **"**OntoKhoj: a Semantic Web portal for ontology searching, ranking and classification"**,** *Proceeding WIDM '03 Proceedings of the 5th ACM international workshop on Web information and data management ACM New York, NY, USA,* 2003.

[33] Segaran, T,. Evans, C,. Taylor, J.,"Programming the Semantic Web", *O'Reilly Media, Inc.*, 84, 2009.

[34] Nolin M.A., Ansell, P., Belleau, F., Idehen, K., Rigault, P., Tourigny, N., Roe,P., Hogan, J.M., Dumontier, M., "Bio2RDF network of linked data", *Semantic Web*

*Challenge; International Semantic Web Conference(ISWC 2008) Karlsruhe, Germany,* 2008.

[35] Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C.S., Willighagen, E., Janos Hajagos, J., Marshall, M.S., Prud'hommeaux, E., Hassenzadeh, O., Pichler, E.& Stephens, S "Linked open drug data for pharmaceutical research and development", *J Cheminform.,* 3(19), 2011.

[36] "Trends in drug discovery and pharmaceutical research", *Clinical Trials,* <http://www.clinical-trials-info.com/drug-discovery-trends/>, 2009.

[37] Hopkins, A.L.,"Network Pharmacology: The Next Paradigm in Drug Discovery", *Nature Chemical Biology*, 4,(11), 2008.

[38] Boran, A.D.W & Iyengar, R., "Systems approaches to polypharmacology and drug discovery ", *Curr Opin Drug Discov Devel*., 13(3): 297–309, 2010.

[39] Druker B.J., Tamura,S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., Zimmermann, J., Lydon, N.B., "Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells", *Nat Med*. 2(5):561-6, 1996.

[40] Yıldırım, M.A., Goh, K.I., Cusick,M.E., Barabási, A.L. & Vidal,M. "Drug–Target Network", *Nature Biotechnology*, 25, (10), 1119, 2007.

[41] Wild D.J., "Mining large heterogeneous datasets in drug discovery**",** *Expert Opinion on Drug Discovery*, 4(10), 995-1004, 2009.

[42] Slater T, Bouton C, Huang E.S. "Beyond data integration", *Drug Discovery Today,* 13(13-14), 584-9, 2008.

[43] Neumann E.K., "A life science Semantic Web: are we there yet?", *Science*, 283,22-5, 2005.

[44] Neumann,E.K., Miller,E, Wilbanks,J. "What the Semantic Web could do for the life sciences". *Drug Discovery Today:BIOSILICO*, 2:228-34, 2006.

[45] Chen, H, Ding, L, Wu, Z, Yu, T, Dhanapalan, L, Chen, J.Y., "Semantic Web for integrated network analysis in biomedicine". *Brief Bioinform,* 10(2):177-92, 2009.

[46] Jentzsch A, Zhao J, Hassanzadeh O, Cheung K, Samwald K, Andersson B, "Linking open drug data". *Proceedings of the International Conference onSemantic Systems (I-SEMANTICS'09), Graz, Austria, 2009*.

[47] Cheung K, Yip K, Smith A, Deknikker R, Masiar A, Gerstein M., "YeastHub: A Semantic Web use case for integrating data in the life sciences domain", *Bioinformatics,* 21(1), i85-96, 2005.

[48] Villanueva-Rosales N, Osbahr K, Doumontier M., "Towards a Semantic Knowledge base for Yeast biologists". *J Biomed Inform.* 41(5),779-89, 2008.

[49] Neumann E.K., Quan D., "Biodash: a Semantic Web dashboard for drug development". *Pac Symp on Biocomput,,* 11,176-187, 2006.

[50] "DrugBank", Univ Alberta, <http://hwww.drugbank.ca>, 2006

[51] "UniProt", European Bioinformatics Institute (EBI)Hinxton U.K. , the Swiss Institute of Bioinformatics (SIB) Geneva, Switzerland, and the Protein Information Resource, Georgetown University Medical Center in Washington, DC, <http://www.uniprot.org>, 2002.

[52] GeneCards, Crown Human Genome Center at the Weizmann Institute of Science, <http://www.genecards.org>, 1997.

[53] National Center for Biotechnology Information (NCBI) of National Institute of Health <http://www.omim.org>, 2004.

[54] Reinefeld, A., & Marsland,T.A., “Enhanced Iterative Deepening Search,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(16), 701-710, 1994.

[55] Russell, S. J., Norvig, P., “Artificial Intelligence: A Modern Approach (2nd ed.)”, Prentice Hall, 2003.

[56] van Rossum, G, "An Introduction to Python for UNIX/C Programmers". *Proceedings of the NLUUG najaarsconferentie (Dutch UNIX users group)*, 1993.

[57] "What is Python Good For?" *General Python FAQ*. Python Software Foundation. Retrieved 09-05-2008.

[58] Forcier, J., Bissex, P., Chun, W.,” Python Web Development with Django (1st ed.)” Addison-Wesley, 408, 2008.

[59] *“Django”,* Django Software Foundation, <http://www.djangoproject.com>*,* 2005.

[60] Allen, G., Owens, M., “The Definitive Guide to SQLite (2nd ed.)”, Apress. 368. 2010.

[61] Wishart D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M., “DrugBank: a knowledgebase for drugs, drug actions and drug targets”, *Nucleic Acids Res: Database issue*. 36, D901-6, 2008.

[62] Wishart D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, “DrugBank: a comprehensive resource for in silico drug discovery and exploration”, *J. Nucleic Acids Res*. *Database issue,* 1(34), D668-72, 2006.

[63] Apweiler, R., Bairoch, A., Wu, C. H. "Protein sequence databases". *Current Opinion in Chemical Biology* 8 (1),76–80, 2004.

[64] Apweiler, R.,Bairoch, A.,Wu, C. H.,Barker, W. C.,Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H. "UniProt: The Universal Protein knowledgebase". *Nucleic Acids Research,* 32 (90001), 115D–1119, 2004

[65] UniProtKB/SwissProt Protein knowledgebase release statistics, 2011

[66] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D., "GeneCards: integrating information about genes, proteins and diseases". *Trends Genet.* 13 (4), 163, 1997.

[67] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D., "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support". *Bioinformatics* 14 (8), 656–64, 1998.

[68] Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Olender T, Chalifa-Caspi V, Lancet D., "GeneCards 2002: towards a complete, object-oriented, human gene compendium". *Bioinformatics,* 18 (11), 1542–3, 2002.

[69] McKusick, V.A. “Mendelian Inheritance in Man; A Catalog of Human Genes and Genetic Disorders”. Baltimore, Maryland: Johns Hopkins University Press., 1998.

[70] Adams, A. R. D., “Drug Treatment of Malaria”, *Br Med J*., 15, 2(5145), 183–184, 1959.

[71] Chong, C.R., Sullivan, D.J., “New uses for old drugs”. *Nature*. 448,645–6, 2007.

[72] Verma, U., Sharma, R., Gupta, P., Kapoor, B., Bano, G., Sawhney, V., “New uses for old drugs: Novel therapeutic options”, *Indian J Pharmacol*, 37(5), 279-287, 2005.

[73] Durk, H.A.,“Maintenance therapy for multiple myeloma with particular emphasis on thalidomide”. *Onkologie,*.29,582–90, 2006.

[74] Jeffreys, D., “Aspirin The Remarkable Story of a Wonder Drug”, New York & London, Bloomsbury, Diane Publishing Company, 2004.

[75] Chan, A.T., Ogino, S., Fuchs, C.S.,“Aspirin and the risk of colorectal cancer in relation to the expression of COX-2”. *N Engl J Med*.,356,2131–42, 2007.

[76] Krumholz, H,M., Radford, M.J., Ellerbeck, E.F, Hennen,J.,. Meehan, T.P., Petrillo, M., Wang, Y., Kresowik, T.F., Jencks, S.F., "Aspirin in the Treatment of Acute Myocardial Infarction in Elderly Medicare Beneficiaries : Patterns of Use and Outcomes". *Circulation* 92 (10): 2841–2847, 1995.

[77] Clive S, Gardiner J, Leonard R.C., “Miltefosine as a topical treatment for cutaneous metastases in breast carcinoma”,*Cancer Chemother Pharmacol*., 44(Suppl), S29–30, 1999.

[78] Sundar, S., Rosenkaimer, F., Makharia, M.K, Goyal, A.K., Mandal, A.K., Voss, A., Hilgard, P., Murray, H.W.. “Trial of oral miltefosine for visceral leishmaniasis”. *Lancet*.,352, 1821–3, 1998.

[79] Webb, D.J., Freestone, S., Allen, M.J., Muirhead, G.J. (March 4,). "Sildenafil citrate and blood-pressure-lowering drugs: results of drug interaction studies with an organic nitrate and a calcium antagonist". *Am. J. Cardiol.* 83 (5A), 21C–28C, 1999.

[80] Boolell, M., Allen, M.J., Ballard, S.A., Gepi-Attee, S., Muirhead, G.J., Naylor, A.M., Osterloh, I.H., Gingell, C., "Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction". *Int J Impot Res* 8 (2), 47–52, 1996.